Comparison of mean square error of prediction or estimation of a full model and a subset model in linear regression

by

Christopher Bingham, R. Dennis Cook, Sanford Weisberg

Technical Report #308

Dept. of Applied Statistics
School of Statistics
University of Minnesota
Saint Paul, MN 55108

1 March 1978

ABSTRACT

Criteria for preferring a subset model over a full linear regression model can be based on comparisons of mean square errors (MSE) of estimation or prediction of the two models. Assuming the full model is unbiased, a subset model of rank > 1 will always have lower MSE for some predictions and may have lower MSE for all predictions. We give the conditions, depending on the eigenstructure of a certain matrix, that determine the pattern of MSE, and discuss a method for estimating MSE averaged over a set of predictions of linear functions of the parameters. This approach leads to Mallows' C_p statistic, as well as to other related statistics. A numerical example is given.

Keywords: Linear regression, subset selection, model building, Mallows' C_p.

1. Introduction

Consider the usual full-rank linear regression model with $p + q$ independent variables (including the constant, if any) which is partitioned to display a selected subset model of interest,

$$Y = \beta_0 J + X\beta + e = \beta_0 J + X_1\beta_1 + X_2\beta_2 + e \qquad (1)$$

where Y is $n \times 1$, X_1 and X_2 are $n \times (p-1)$ and $n \times q$, respectively, J is the $n \times 1$ vector of 1's and $\beta = (\beta_1', \beta_2')'$ is $((p-1) + q) \times 1$. For convenience, we assume that the $p-1+q$ columns of X are orthogonal to J but not necessarily to each other. An important problem in linear regression is the comparison of the full model (1) to the subset model

$$Y = \beta_0 J + X_1\tilde{\beta}_1 + \tilde{e} \qquad (2)$$

Models (1) and (2) are equivalent if $\beta_2 = 0$. A subset model would be preferred only if it were in some sense better, perhaps providing more precise predictions or fitted values for some region of interest. (See Hocking, 1976, for a review of standard variable selection criteria and methodology.) Although (2) is, in general, not unbiased, it may be preferable to (1) for specific purposes. It is well known, for example, that (2) may have smaller mean square error for estimation of parameters.

In this article, using a mean square error criterion, we extend the standard methodology for comparing (1) and (2) under the assumption that the full model is unbiased and has homoskadastic uncorrelated errors, that is $E(e) = 0$ and $Cov(e) = \sigma^2 I_n$. This

criterion leads to examination of a new function of the data and gives, as special cases, Mallows' C_p statistics (Mallows, 1973), and the subdivision of C_p suggested by Weisberg (1977).

2. Comparing a fixed subset model to the full model

Let $\hat{\beta}_F = S^{-1}X'Y$ denote the least squares estimator of β from the full model (1), where

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix},$$

and let

$$\hat{\beta}_S = \begin{pmatrix} \hat{\tilde{\beta}}_1 \\ 0 \end{pmatrix} = \begin{pmatrix} S_{11}^{-1}X_1'Y \\ 0 \end{pmatrix}$$

be the least squares estimator of β under (2). Under our assumptions $\hat{\beta}_0 = \bar{Y}$ for both (1) and (2). Now, suppose z' is any $1 \times (p-1+q)$ vector, in the row space of X. In particular z' could be a row of X or z' might represent a contrast or linear function of one or more of the elements of β. We wish to compare the mean square errors of estimates of $z'\beta$, under the two models (1) and (2). These estimates are $z'\hat{\beta}_S$ and $z'\hat{\beta}_F$, respectively. A comparison of mean square error of prediction will yield the same result as a comparison of mean square error of estimation. Moreover, a comparison of $\bar{Y} + z'\hat{\beta}_F$ to $\bar{Y} + z'\hat{\beta}_S$ will lead to the same results.

For any estimator, $\hat{\beta}$, the mean square error of $z'\hat{\beta}$ is

$$MSE\,(z'\hat{\beta}) = E(z'\hat{\beta}-z'\beta)^2 = z'Var(\hat{\beta})z + z'(\beta-E(\hat{\beta}))'(\beta-E(\hat{\beta}))'z.$$

Let $D(z)$ be defined by

$$D(z) \equiv MSE\,(z'\hat{\beta}_S) - MSE(z'\hat{\beta}_F) \ . \qquad (3)$$

We shall say that the subset model is <u>better</u> than the full model <u>at z</u> if $D(z) < 0$; if $D(z) > 0$ the full model is better (if $D(z) = 0$, the models are equivalent). If $D(z) < 0$ for every z' in the row space of X we say the subset model is <u>everywhere better</u>.

Given models (1) and (2) and $Cov(e) = \sigma^2 I$, it is straightforward to find $D(z)$:

$$D(z) = z'(P'Q\,P)z \quad , \qquad (4)$$

where $P = [-S_{21}S_{11}^{-1},\ I_q]$, $\quad Q = \beta_2\beta_2'/\sigma^2 - S_{22\cdot 1}^{-1}$,

and $\quad S_{22\cdot 1} = S_{22} - S_{21}S_{11}^{-1}S_{12}$.

The matrices, P, β_2, and $S_{22\cdot 1}$ all reflect the role of the variables to be omitted in (2) after adjusting for the variables included in both models.

<u>Result 1</u> (Hocking, 1974). $D(z) < 0$ for all z if and only if Q is negative definite ($Q < 0$) or, equivalently, if and only if the eigenvalues of Q are all negative.

This result implies that, if $Q < 0$, the subset model is everywhere better. As a particular case, suppose $z' = (0', z_2')$ and thus $z'\beta$ is a linear function of elements of β_2. If $Q < 0$, the estimator $z'\hat{\beta}_S = 0$ has lower mean square error than the estimator from the full model.

Although the eigenvalues and eigenvectors of Q are not invariant under linear transformations of X, it is easily seen that

the signs of the largest and smallest eigenvalues are invariant. In particular, we can transform $X = [X_1 \; X_2]$ into $X^* = [X_1 \; X_2^*]$ by

$$X^* = [X_1 \; X_2] \begin{bmatrix} I_p & -S_{11}^{-1}S_{12}S_{22\cdot 1}^{-1/2} \\ 0 & S_{22\cdot 1}^{-1/2} \end{bmatrix}.$$

Then,

$$X^{*\prime}X^* = \begin{bmatrix} S_{11} & 0 \\ 0 & I_q \end{bmatrix}.$$

In terms of X^*, model (1) becomes

$$Y = \beta_0 J + X_1\alpha_1 + X_2^*\alpha_2 + e$$

where $\alpha_1 = \beta_1 + S_{11}^{-1}S_{12}\beta_2$, $\alpha_2 = S_{22\cdot 1}^{1/2}\beta_2$. For this semi-orthogonal model, the matrix Q becomes

$$Q^* = \alpha_2\alpha_2'/\sigma^2 - I_q = S_{22\cdot 1}^{1/2} Q S_{22\cdot 1}^{1/2} . \qquad (5)$$

Since $\alpha_2\alpha_2'$ is of rank 1, the eigenvalues of Q^* are -1 with multiplicity $q-1$, and $\alpha_2'\alpha_2/\sigma^2 - 1$ with multiplicity 1. Thus at least $q-1$ eigenvalues are negative, with the remaining one being positive if and only if $\alpha_2'\alpha_2/\sigma^2 = \beta_2' S_{22\cdot 1} \beta_2 > 1$. We sum this up as

Result 2. All eigenvalues of Q are negative and hence the subset model is everywhere better if and only if $\lambda = \beta_2' S_{22\cdot 1} \beta_2 < 1$ (Hocking, 1976). Moreover, if $q > 1$, Q always has at least one negative eigenvalue, and there always exists a linear subspace of the row space of X such that for any z' in that subspace, $D(z) < 0$ and the subset model is better.

Generally Q will have eigenvalues of both signs, and neither the subset model nor the full model will be better for all z. Note that λ is the noncentrality parameter for the usual F test of $H_0: \beta_2 = 0$.

3. Estimation of Q and $D(z)$

In practice $\beta_2\beta_2'$ is unknown, and hence neither Q nor $D(z)$ can be known exactly. A familiar approach is to estimate Q by some function of data, say $\hat{Q}$, and then estimate $D(z)$ by replacing Q by $\hat{Q}$ in (4). Writing $\hat{\beta}_F = (\hat{\beta}_1', \hat{\beta}_2')'$ and using the assumption that $\mathrm{Cov}(e) = \sigma^2 I$ for model (1), it follows that

$$E(\hat{\beta}_2\hat{\beta}_2'/\sigma^2 - 2S_{22.1}^{-1}) = Q .$$

This suggests defining $\hat{Q}$ as

$$\hat{Q} = \hat{\beta}_2\hat{\beta}_2'/\hat{\sigma}^2 - 2S_{22.1}^{-1} , \tag{6}$$

where $\hat{\sigma}^2$ is an estimator of σ^2. As usual, we take $\hat{\sigma}^2 = s^2$, the residual mean square from the full model (1). The estimator of $D(z)$ is then

$$\hat{D}(z) = z'(P'\hat{Q}P)z. \tag{7}$$

This estimator of the normalized mean square error at z is equivalent to one proposed by Allen (1971), where it was proposed as a criterion for model selection based on mean square error of prediction at a single point.

In terms of the orthogonalized coordinates X^*, the estimator of Q is

$$\hat{Q}^* = \hat{\alpha}_2\hat{\alpha}_2'/s^2 - 2I_q , \quad \hat{\alpha}_2 = S_{22\cdot 1}^{1/2}\hat{\beta}_2 = X_2^{*\prime}Y . \tag{8}$$

The eigenvalues of $\hat{Q}^*$ are -2 with multiplicity $q-1$ and $\hat{\alpha}_2'\hat{\alpha}_2/s^2 - 2$. But

$$\hat{\alpha}_2'\hat{\alpha}_2/s^2 = \hat{\beta}_2'S_{22\cdot 1}\hat{\beta}_2/s^2 = qF_S ,$$

where F_S is the usual normal theory test statistic of $H_0:\beta_2 = 0$. Thus, the largest eigenvalue of $\hat{Q}$ will be negative if and only if $F_S < 2/q$. This is a very strong requirement that seldom will be satisfied for $q > 2$, even if $\beta_2 = 0$. Assuming normality, F_S is distributed as a non-central $F(q, n-p-q; \lambda)$, with noncentrality parameter $\lambda = \beta_2' S_{22\cdot 1}\beta_2/\sigma^2$. Tests of the null hypothesis that Q is negative definite can be based on this distribution with $\lambda = 1$.

When $q = 1$, $\hat{Q} = \hat{\beta}_2^2/s^2 - 2S_{22\cdot 1}^{-1} = (t^2-2)S_{22\cdot 1}^{-1}$ where t^2 is the usual likelihood ratio test statistic (under normality) of $\beta_2 = 0$. In this case $\hat{D}(z)$ will be negative for all z if $t^2 < 2$, and positive for all z if $t^2 > 2$.

4. Mean square error over a region

An important extension of the results so far can be obtained as follows. Let $Z = [Z_1\ Z_2]$ be an $m \times (p-1+q)$ matrix whose rows determine a set of linear functionals of interest and define the $m \times m$ matrix

$$\hat{D}(Z) = Z(P'\hat{Q}\ P)Z' \qquad (9)$$

In particular, if the set of linear functionals is the subspace spanned by the rows of Z the subset model will appear to be better throughout the subspace if $\hat{D}(Z)$ is negative semi-definite. Such a subspace always exists if $q \geq 2$ since the eigenvectors of $P'\hat{Q}\ P$ corresponding to negative eigenvalues will span a subspace in which the subset model is estimated to have smaller MSE. For this reason, in practice the computation of the non-zero eigenvalues of $P'\hat{Q}\ P$ will be more useful than the computation of the eigenvalues of $\hat{Q}$.

One interesting set of linear functionals of importance for comparing model (2) to model (1) is the rows of X itself. This suggests examining certain functions of $\hat{D}(X)$, such as the signs of its eigenvalues, its trace, diagonal elements, maximum diagonal element, and so on.

Mallows' C_p. The C_p statistic is defined to be $C_p = RSS_p/s^2 + 2p - n$ where RSS_p is the residual sum of squares in the subset model (Daniel and Wood 1971). Using the cyclic permutation invariance of the trace and $PX'XP' = S_{22\cdot 1}$,

$$\text{tr } \hat{D}(X) = \hat{\beta}_2' PX'XP' \hat{\beta}_2/s^2 - 2\text{tr } PX'XP' S_{22\cdot 1}^{-1}$$

$$= (RSS_p - (n-p-q)s^2)/s^2 - 2q \ .$$

Thus we have

Result 3:

$$C_p = \text{tr } \hat{D}(X) + (p+q)$$

where $p + q$, the total number of parameters in the full model, is fixed.

Mallows' C_p essentially averages mean square error over all the rows in the data. A model with relatively low C_p may have both large negative and large positive eigenvalues of $\hat{Q}$, which average to give a small value of C_p. However, if $\hat{Q}$ is negative definite, that is, if the subset model is estimated to be better, then $C_p - (p+q) = \text{tr } (\hat{D}(X)) < 0$, so a better model must have $C_p < (p+q)$ although this is not sufficient.

If we define F_S as before, it follows easily that

$$\text{tr } (\hat{D}(X)) = q(F_S - 2)$$

which suggests comparing F_S to 2, a much less restrictive condition than comparing it to $2/q$. Even less restrictive is to compare F to the

upper α x 100% point of $F(q,n-p-q;1)$ which will be exceeded only if there is is statistical evidence the subset model is not better.

The diagonal elements $\hat{D}(X)_{ii}$, of $\hat{D}(X)$, also have interest in their own right. The subset model is estimated to be better for the i-th individual if and only if $\hat{D}(X)_{ii} < 0$. These statistics are also related to a statistic given in an unpublished paper by Weisberg (1977). He defined a partition of Mallows' C_p into n pieces, one for each individual. These statistics are defined by

$$C_{pi} = \frac{(x_i'(\hat{\beta}_F-\hat{\beta}_S))^2}{s^2} + 2v_i - u_i \quad , \quad i=1,2,\ldots n,$$

where $x_i' = (x_{i1}', x_{i2}')$ is the i-th row of X, $u_i = x_i' S^{-1} x_i$ is the variance of $x_i'\hat{\beta}_F/\sigma$, and $v_i = x_{i1}' S_{11}^{-1} x_{i1}$ is the variance of $x_i'\hat{\beta}_S/\sigma$. The C_{pi} are defined so that $\Sigma C_{pi} = C_p$.

The C_{pi}, and the $\hat{D}(X)_{ii}$ depend on Y only through the sufficient statistics $Y'Y$, $J'Y$, and $X'Y$. Thus they are not influenced by outliers, except insofar as such outliers affect the least squares estimators.

Also, let $t_i^2 = [x_i'(\hat{\beta}_F-\hat{\beta}_S)]^2/s^2(u_i-v_i)$ be the normal theory likelihood ratio test statistic of no bias for the i-th individual. Then one can rewrite

$$\hat{D}(X)_{ii} = (u_i - v_i)(t_i^2-2) .$$

If $\hat{Q}$ is negative definite then $\hat{D}(X) < 0$, $tr(\hat{D}(X)) < 0$ and $t_i^2 < 2$ for all x_i. For any $\hat{Q}$, the $\hat{D}(X)_{ii}$ satisfy

$$-2(u_i - v_i) \leq \hat{D}(X)_{ii} \leq tr\, \hat{D}(X) + 2(q-(u_i - v_i)) .$$

The $\hat{D}(X)_{ii}$ are essentially compounded of a fixed quantity $(u_i - v_i)$, and a random quantity t_i^2. The former quantity measures in a general way the impact of the i-th row on the regression estimates (see Hoaglin and

Welsch (1978) for a discussion). Large values of u_i correspond to potentially important points in the data and the difference $u_i - v_i$ measures the change in importance of the i-th individual when changing from the full model to the subset model. If $u_i - v_i$ is near zero, then for the i-th individual the deleted variables are nearly equal to the means of those variables (or are nearly zero if the mean is not included in the subset). Consequently, these additional variables carry little information for this individual, and only negligible bias should result (for this individual) if the subset model is used, even if $t_i^2 > 2$. On the other hand, if $u_i - v_i$ is large, then a large potential for bias exists since the deleted variables may be important. By the same argument, if $u_i - v_i$ is large, the potential savings in mean square error for this individual is also large when the bias is small.

5. Example.

As an example we examine the well known data on the hardening of cement first given by Hald (1952), and later analyzed in detail by Draper and Smith (1966), Daniel and Wood (1971), and Seber (1977). The data is given on pages 365-66 of Draper and Smith (1966), and in the other references. In this data, $n = 13$, and $p + q = 5 = 1 +$ number of independent variables (labeled X_1, X_2, X_3, X_4).

The non-zero eigenvalues and corresponding eigenvectors of $P' \hat{Q} P$ for several subset models are given in Table 1. The signs of the smallest and largest eigenvalues of this matrix are the same as the signs of the smallest and largest eigenvalues of $\hat{Q}$. The first model given in Table 1, (X_1, X_2), minimizes C_p. Since both eigenvalues are

negative we can conclude that this model is estimated to be everywhere better than the full model. The improvement over the full model is greatest for points in the direction of the eigenvector corresponding to the largest negative eigenvalue (i.e., all differences $X_j - \bar{X}_j$ approximately equal, j=1,2,3,4) and smallest in the direction of the other eigenvector (i.e., $X_1 - \bar{X}_1 \cong X_3 - \bar{X}_3 \cong -(X_2 - \bar{X}_2) \cong -(X_4 - \bar{X}_4)$). The nominal p-value for a test of $H_0:\lambda = 1$ based on non-central $F(2,8;1) = .8391$ is $p = .608$. Thus there is no statistical evidence that the full model is better than (X_1, X_2) at any point in the independent variable space.

Tables 1 and 2 about here

The model (X_1, X_4) has eigenvalues of each sign, so neither the full model nor the subset model is estimated to always have lower mean square error. The p-value for $F(1,8;1) = 2.2482$ is .285. In Table 2 are the $\hat{D}(X)_{ii}$ for this and other models. For (X_1,X_2) all the $\hat{D}(X)_{ii}$ are negative and for (X_1,X_4), 7 of 13 of the $\hat{D}(X)_{ii}$ are positive, indicating individuals for which the full model appears to be better than (X_1,X_4). For individual 5 the full model appears substantially better. This seems to indicate that the model (X_1,X_2) may be more appropriate than (X_1,X_4), a conclusion that differs from previous claims (e.g., Seber, 1977, p. 360).

Both of the three variable models that include (X_1,X_2) are estimated to be everywhere better than the full model. Tables 1 and 2 show that these two models are virtually indistinguishable with respect to estimated mean square error.

If a single, final model were desired, one would choose between either (X_1, X_2) or one of the three variable models discussed. On the basis of the analysis here, there is little to differentiate between them.

The fifth model in Tables 1 and 2, X_3 alone, is included because it has the largest C_p of any of the 15 possible models, $C_p = 315.15$. Even for this very poor model, at least one of the eigenvalues must be negative (actually, two of them are negative) and, for points in the direction of the corresponding eigenvectors, the subset model will have lower estimated MSE; this occurs for individual 11. However, for the remaining individuals, this model is substantially worse than the full model.

6. Discussion.

In a selection problem, a reasonable approach is to limit consideration to subsets with relatively small values of C_p (or tr $\hat{D}(X)$), certainly considering only models with tr $\hat{D}(X) \leq 0$, or, equivalently, $C_p \leq p + q$. For these models, the eigenvalues and eigenvectors of $\hat{Q}$ or $P'\hat{Q}\,P$ should be computed. Examination of these statistics may give further guidance in selecting variables. If the goal of the problem is prediction or estimation of fitted values, a good subset model should have lower mean square error of prediction for all or nearly all predictions in a region of interest. The eigenvectors of $P'\hat{Q}\,P$ corresponding to negative eigenvalues define the subspace over which the submodel is estimated to be better.

Finally, if the full model itself is biased then the results given here may no longer be valid. In particular, if independent variables with

nonzero coefficients have been left out of the full model, then it can be shown that $D(z)$ always has at least one positive and at least one negative eigenvalue, and there always exists a linear subspace where the subset model is worse than the full model.

References

Allen, D.M. (1971). Mean square error of prediction as a criterion for selecting variables. Technometrics 13, 469-75.

Daniel, C. and Wood, F. (1971). Fitting Equations to Data. New York: Wiley.

Draper, N. and Smith, H. (1966). Applied Regression Analysis. New York: Wiley.

Hald, A. (1952). Statistical Theory and Engineering Applications. New York: Wiley.

Hoaglin D, and Welsch, R. (1978). The hat matrix in regression and anova. American Statistician 32, 17-22.

Hocking, R.R. (1974). Misspecification in regression. The American Statistician 28, 39-40.

Hocking, R.R. (1976). The analysis and selection of variables in linear regression. Biometrics 32, 1-49.

Mallows, C.L. (1973). Some comments on C_p. Technometrics 15, 661-75.

Seber, G.A.F. (1977). Linear Regression Analysis. New York: Wiley.

Weisberg, S. (1977). A statistic for allocating C_p to individual cases. University of Minnesota, School of Statistics Technical report No. 296.

Table 1: Non-zero eigenvalues and corresponding eigenvectors of $P'\hat{Q}\,P$ for 5 models.

model	C_p	eigen-value	eigenvectors			
			X_1	X_2	X_3	X_4
$X_1\ X_2$	2.68	-0.7011 -0.0019	.505 .446	.492 -.526	.518 .528	.484 -.495
$X_1\ X_4$	5.50	-0.6455 0.0165	.530 .436	.456 -.493	.461 -.581	.546 .479
$X_1\ X_2\ \ X_4$	3.02	-0.6833	.511	.485	.477	.525
$X_1\ X_2\ X_3$	3.04	-0.6756	.496	.503	.507	.494
X_3	315.15	-0.4293 -0.0024 0.4739	.282 .038 -.573	.203 -.354 -.704	.551 .792 -.025	.759 -.495 .419

Table 2: $\hat{D}(X)_{ii}$ for 5 models

i	X_1X_2	X_1X_4	$X_1X_2X_4$	$X_1X_2X_3$	X_3
1	-.18	.39	-.059	-.18	96.85
2	-.11	-.11	-.11	-.077	57.21
3	-.91	-.82	-.88	-.89	4.71
4	-.11	-.041	-.10	-.10	19.22
5	-.097	.69	-.0005	-.045	7.81
6	-.016	-.011	-.013	-.016	6.73
7	-.007	.010	-.007	-.005	38.63
8	-.13	.17	-.12	-.050	7.72
9	-.20	-.11	-.17	-.21	2.56
10	-.11	.17	-.095	-.032	17.21
11	-.22	.079	-.21	-.092	- .14
12	-.13	-.063	-.12	-.13	29.79
13	-.12	.15	-.088	-.13	21.86